Posterior Predictive Densities, Posterior Predictive p-values, Highest Posterior Density Interval

(KPT, Ch. 7)
R scripts: `mod6_4a, mod6_4b, mod6_4c`

Posterior Predictive Densities

Consider again the female wage regression from script `mod5_3e`. Let's assume you're interested in predicting the outcome (in this case the log(earnings), and – ultimately – earnings in $$$) for a woman with characteristics $\mathbf{x_p}$, i.e. you are interested in learning about the construct

$$\hat{\mathbf{y}}_\mathbf{p} = \mathbf{x}_\mathbf{p}'\boldsymbol{\beta} + \varepsilon_p \tag{1}$$

In Bayesian analysis this means you're interested in the *posterior predictive distribution* of $\hat{y}_p$. Formally, this PPD is given by (suppressing conditionality on $\mathbf{x_p}$ for convenience)

$$\begin{aligned} p(\hat{y}_p \mid \mathbf{y}) &= \int_{\boldsymbol{\theta}} p(\hat{y}_p, \boldsymbol{\theta} \mid \mathbf{y})\, d\,\boldsymbol{\theta} = \int_{\boldsymbol{\theta}} p(\hat{y}_p \mid \boldsymbol{\theta}, \mathbf{y})\, p(\boldsymbol{\theta} \mid \mathbf{y})\, d\,\boldsymbol{\theta} = \int_{\boldsymbol{\theta}} p(\hat{y}_p \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathbf{y})\, d\,\boldsymbol{\theta} \\ \boldsymbol{\theta} &= \left[\boldsymbol{\beta}' \quad \sigma^2\right]' \end{aligned} \tag{2}$$

where the third equality follows from the fact that, conditional on $\boldsymbol{\theta}$, $\hat{y}_p$ is independent from the actual data $\mathbf{y}$. From our basic regression framework we know that

$$p(\hat{y}_p \mid \boldsymbol{\theta}) = n\left(\mathbf{x}_\mathbf{p}'\boldsymbol{\beta}, \sigma^2\right) \tag{3}$$

Naturally, $p(\boldsymbol{\theta} \mid \mathbf{y})$ is simply our posterior distribution, and we already have draws from it from our original Gibbs Sampler. Thus, we can derive $p(\hat{y}_p \mid \mathbf{y})$ by drawing $\hat{y}_p$ from $n\left(\mathbf{x}_\mathbf{p}'\boldsymbol{\beta}_r, \sigma_r^2\right)$, where subscript *r* indicates the r^{th} draw of our parameters in the retained series of draws from the original GS. To be precise, 1 draw of $\hat{y}_p$ per $\boldsymbol{\theta}_r$ is sufficient to generate the PPD. Optionally, you can take several draws of $\hat{y}_p$ per $\boldsymbol{\theta}_r$ for a "smoother" posterior density (= nicer to plot).

Once you have all draws of $\hat{y}_p$ (say "*R*" of them), you can derive the moments of the resulting PPD via Monte Carlo integration. For example, the posterior predictive mean can be derived as

$$E(\hat{y}_p \mid \mathbf{y}) = \int_{\hat{y}_p} \hat{y}_p\, p(\hat{y}_p \mid \mathbf{y})\, d\,\hat{y}_p \approx \frac{1}{R}\sum_{r=1}^{R} \hat{y}_{pr} \tag{4}$$

In regression analysis, we are often interested in predicting a function of $\hat{y}_p$. The most prominent example is $\exp(\hat{y}_p)$ when the original regression used the log form for the dependent variable. In the Bayesian context, this implies that we're interested in the PPD of $\exp(\hat{y}_p) \mid \mathbf{y}$. You can think of this in one of two ways – the hard way and the easy way.

The hard way would be to re-define your posterior construct of interest as

$$\tilde{y}_p = \exp\left(x_p'\boldsymbol{\beta} + \varepsilon_p\right) \tag{5}$$

where I use the "~" symbol to distinguish the predictive construct to our original one in (1). Formally, we have again

$$p\left(\tilde{y}_p \mid \mathbf{y}\right) = \int_{\boldsymbol{\theta}} p\left(\tilde{y}_p \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta} \mid \mathbf{y}\right) d\,\boldsymbol{\theta} \tag{6}$$

In theory, we could proceed as before to generate draws from the PPD. However, this requires drawing from $p\left(\tilde{y}_p \mid \boldsymbol{\theta}\right)$ - and it's not immediately clear what the analytical form of this density looks like (in this case it's a log-normal, in fact, but in many cases the form of $p\left(\tilde{y}_p \mid \boldsymbol{\theta}\right)$ will be unknown or difficult to assess).

The easy way would be to realize that you already have draws of $\hat{y}_p = \log\left(\tilde{y}_p\right)$ (or you know how to generate them easily following the procedure form above). Then by exponentiating each draw of $\hat{y}_p$ we automatically get $p\left(\tilde{y}_p \mid \mathbf{y}\right) = p\left(\exp\left(\hat{y}_p\right) \mid \mathbf{y}\right)$. Similarly, the moments of $\tilde{y}_p$ can be approximated via Monte Carlo Integration, for example:

$$\begin{aligned} E\left(\tilde{y}_p \mid \mathbf{y}\right) &= \int_{\tilde{y}_p} \tilde{y}_p\, p\left(\tilde{y}_p \mid \mathbf{y}\right) d\,\tilde{y}_p \approx \frac{1}{R}\sum_{r=1}^{R} \tilde{y}_{pr} = \\ E\left(\exp\left(\hat{y}_p\right) \mid \mathbf{y}\right) &= \int_{\hat{y}_p} \exp\left(\hat{y}_p\right) p\left(\hat{y}_p \mid \mathbf{y}\right) d\,\hat{y}_p \approx \frac{1}{R}\sum_{r=1}^{R} \exp\left(\hat{y}_{pr}\right) \end{aligned} \tag{7}$$

R script `mod6_4a` illustrates this procedures.

Posterior Predictive P-values (PPPs)

The computation of PPPs builds directly on predictive distributions, so this is a good place to discuss them.

PPPs are an alternative measure of "model fit", especially for cases where the marginal likelihood is difficult to derive. They can be used to assess the qualities of a single model, or to compare several models (in which case a model with higher PPP would be preferred). PPPs can also be used to test some of the assumptions underlying a given model.

The basic intuition for PPPs (or "Bayes p-values") is as follows: We choose some *construct of interest or test statistic* of interest that is a function of our actual data $\mathbf{y}$. We then simulate a posterior predictive density for the same statistic, *using simulated or "predicted" data instead of* ***y***. We then examine where in this distribution the original statistic (the one that depends on actual data **y**) is located. If it's far out in the tail, we conclude that it is very unlikely that our observed data could have been generated by our

specified modeling structure. "Far out" means that the area that's left in the tail, i.e. the p-value, is small. As a rule of thumb, a p-value of 0.05 or smaller would be interpreted as evidence against a given model.

In most cases, our construct of interest will be a function of both **y** and parameters $\boldsymbol{\theta}$. Call it $T(\mathbf{y},\boldsymbol{\theta})$. A popular example in regression analysis is the Skewness statistic (*sk*), given as

$$sk = \frac{\sqrt{N}\sum_{i=1}^{n}\varepsilon_i^3}{\left[\sum_{i=1}^{n}\varepsilon_i^2\right]^{\frac{3}{2}}} \quad \text{where} \quad \varepsilon_i = y_i - \mathbf{x_i'}\boldsymbol{\beta} \tag{8}$$

If our original assumption of normality for regression errors is correct, we would expect *sk* to lie close to zero.

To test this assumption (and thus the appropriateness of our model) we start by generating a posterior predictive distribution of $sk \mid \mathbf{y}$ following exactly the steps outlined above for PPDs: For each draw of $\boldsymbol{\beta}_\mathbf{r}$ from the original Gibbs Sampler, compute sk_r using (8) with actual data **y** and **X**. The result is the posterior predictive density of *sk*, i.e

$$p(sk \mid \mathbf{y}) = \int_{\boldsymbol{\theta}} p(sk \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathbf{y})\, d\,\boldsymbol{\theta} \tag{9}$$

We now repeat this process, but now, for each $\boldsymbol{\beta}_\mathbf{r}$ and σ_r^2, we first draw *N* observations of from $n(\mathbf{x_i'}\boldsymbol{\beta}, \sigma^2), i = 1...N$. (*N* is the original sample size). That's our "simulated data" for round *r*. Call the resulting vector $\mathbf{y}_\mathbf{r}^*$. Use this to compute the skewness sk_r^*. Repeat this process *R* times. The result is the posterior predictive distribution of *sk* based on simulated data, i.e.

$$p(sk^* \mid \mathbf{y}) = p(sk(\mathbf{y}^*) \mid \mathbf{y}) = \int_{\boldsymbol{\theta}} p(sk(\mathbf{y}^*) \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathbf{y})\, d\,\boldsymbol{\theta} \tag{10}$$

We now have <u>two options</u> to assess if our regression model, with its normality assumption, is suitable for our actual data. First, we can take all R draws from (9) and (10), and compute the difference $(sk_r^* - sk_r), r = 1...R$. The resulting distribution of differences should be centered and tight around zero if our model is correct (i.e. fits the underlying data well, as judged by the skewness criterion).

The second and more formal approach is via PPPs. This requires a point estimate of $sk|\mathbf{y}$, and we generally use the mean, i.e.

$$E(sk \mid \mathbf{y}) = \int_{sk} sk \cdot p(sk \mid \mathbf{y})\, d\, sk \approx \frac{1}{R}\sum_{r=1}^{R} sk_r \tag{11}$$

We now plot the density of *sk** and examine where $E(sk \mid \mathbf{y})$ is located relative to this distribution. If it's "far out" in one of the tails, our model is unlike to fit the observed data. Formally, we can compute a numerical PPP value as

$$PPP = \min(x, 1-x) \qquad \text{where} \quad x = prob\left(\left(sk^* \mid \mathbf{y}\right) > E\left(sk \mid \mathbf{y}\right)\right) \tag{12}$$

R script `mod6_4b` provides an example.

Highest Posterior Density Intervals (HPDIs)

HDPIs are another tool for examining posterior outcomes and for model comparison when the marginal likelihood is difficult to compute or when the model uses improper priors, in which case a model comparison based on marginal likelihood may not be permissible.

If used for model comparison HDPIs are geared towards the comparison of *nested* models. A *nested* model is one that can be derived from an unconstrained (full) model by imposing linear restrictions on some of the parameters. The most common case is setting one of the slope coefficients in a regression model to zero (i.e. "dropping a regressor").

We start with the definition of a "credible set" C for a parameter, a set of parameters, or a function of parameters. Suppose that a parameter vector $\boldsymbol{\beta}$ of length *k* can take any real value for all its elements, i.e. $\boldsymbol{\beta} \in R^k$. Let $\omega = g(\boldsymbol{\beta})$ be some *m*-vector of functions of $\boldsymbol{\beta}$, defined over a region of $\boldsymbol{\Omega}$, with $m \le k$. Let *C* be a region within Ω, i.e $C \subseteq \Omega$. Assume you have derived the posterior density of $p(\omega \mid \mathbf{y})$. Then a $100(1-\alpha)\%$ credible set with respect to $p(\omega \mid \mathbf{y})$ is given by

$$p(\omega \in C \mid \mathbf{y}) = \int_C p(\omega \mid \mathbf{y})\, d\omega = 1-\alpha \tag{13}$$

The simplest example is the credible set for a single parameter, also called "*credible interval*". Consider parameter β with posterior density $p(\beta \mid \mathbf{y})$. So in this case $\omega = g(\boldsymbol{\beta}) = \beta$. Then *C* is the interval [a b], such that β falls within these bounds with probability $(1-\alpha)$. As for classical confidence intervals, we usually set $\alpha = 0.05$. Thus:

$$p(a \le \beta \le b \mid \mathbf{y}) = \int_a^b p(\beta \mid \mathbf{y})\, d\beta = 0.95 \tag{14}$$

Typically, there are numerous such intervals for a given parameter. We usually choose the one with the smallest area . This makes the credible interval the "*highest posterior density interval*" (HPDI). HPDIs are often reported along with posterior moments and convergence diagnostics as part of the posterior output.

So the first and foremost purpose of a HPDI is to find the bounds *a* and *b* for a specific parameter, such that we can be $100(1-\alpha)\%$ sure that the parameter lies between them. If $p(\beta \mid \mathbf{y})$ is unimodal and symmetric the 95% HPDI is simply the interval between the 2.5th and the 97.5th percentile, what Gelman et al (section 2.3) refer to as "central posterior interval". These percentiles are usually known for common densities (e.g. -1.96 to 1.96 for the standard normal), or can be easily computed analytically. However, the exact form of $p(\beta \mid \mathbf{y})$ is usually unknown. This requires the derivation of the HPDI via simulation and empirical frequencies.

Second, the HPDI can be used as an *ad hoc* method (i.e. a method that is not firmly rooted in probability theory) to test linear model restrictions. For example one might consider a model for which $\beta = c$ (some constant, often zero). If the HPDI for β does not include *c*, this would provide informal evidence against the constrained model. Similarly, for any scalar-valued linear restriction involving several parameters, say $\mathbf{R}\boldsymbol{\beta} = c$, we can first derive the posterior density $p(\mathbf{R}\boldsymbol{\beta} \mid \mathbf{y})$, and examine if the corresponding HPDI includes *c*. See script `mod6_4c` for an example of HDPI computation.